Computer Arithmetic - Spring 2000
Assignment No. 3
Course homepage: http://www.eng.tau.ac.il/~guy/arith00/

Firm Deadline: Sept. 15th, 2000 (in my mail box in Room 124).

Reminder: Projects are due by Sept. 30, 2000.

Questions: (submit answers to 4 questions.)

1. Floating-point addition. Why should the alignment shift be limited by $p+2$? Give an example where limiting the alignment shift by $p+1$ yields a wrong answer.

2. Floating-point addition. Let x_1 and x_2 denote two representable floating point numbers (say, in double precision). Let $x = x_1 + x_2$. Prove that, for every rounding mode, if $abs(x) < 2^{e_{\min}}$, then $r(x) = x$. (Assume that rounding is done also in double precision.)

3. Prove that $r_{32}(r_{64}(x)) \neq r_{32}(x)$ in round-to-nearest (even) mode, where r_{64} denotes round to double precision and r_{32} denotes round to single precision. Assume that $2^{e_{\min}} \leq x \leq 2^{e_{\max}}$ (so overflow and underflow do not occur).

4. Assume the following situation: the arguments of a floating point operation are double precision but the result should be given in single precision. The implementation is to (a) first compute $r_{64}(x)$ (assuming that the result is in double precision); and (b) if the result is required in single precision, then it is "fixed". Suggest a method to compute $r_{32}(x)$ correctly from $r_{64}(x)$ using only 2 additional bits that are computed during the computation of $r_{64}(x)$.

5. Show how to compute the sticky bit of a carry-save encoded binary string. Your construction should be not worse than the following: a constant depth preprocessing followed by an OR-tree.

   Input: $c[1:q-1], s[1:q-1] \in \{0,1\}^q$.
   Output: $sticky\_bit(f,0)$, where $f[0:q-1]$ satisfies

   $$\langle f[0:q-1]\rangle = \langle c[1:q-1]\rangle + \langle s[1:q-1]\rangle.$$

   (Note that all the strings represent fractions.)

   Hints:

   (a) Show that feeding an OR-Tree with the bit-wise XOR of $c[1:q-1]$ and $s[1:q-1]$ does not compute $sticky\_bit(f,0)$.

   (b) Consider the recoding

   $$\langle c[1:q-1]\rangle + \langle s[1:q-1]\rangle = \langle x[0:q-2]\rangle - \langle y[1:q-1]\rangle$$

   obtained by ppm-addition:

   $$(x,y) = ppm(c,s,0).$$

   Let $g[i] = x[i] - y[i]$ (define $x[q-1] = y[0] = 0$). Prove that the fraction range of $g[0:q-1]$ equals $(-1/2, 1)$.

   (c) Prove that $sticky\_bit(f,0)$ equals the $OR(h[1],\ldots,h[q-1])$, where $h[i] = XOR(x[i], y[i])$.